Ondokuz Mayıs Üniversitesi Eğitim Fakültesi Dergisi
Ondokuz Mayıs University Journal of Faculty of Education
e-ISSN: 2548-0278 OMU EFD, December 2024, 43(2): 951-970

Investigation of Models Used in Equating Testlet-Based Tests

Madde Takımı Tabanlı Testlerin Eşitlenmesinde Kullanılan Modellerin İncelenmesi

Ertunç UKŞUL[1], **Hülya KELECİOĞLU**[2]

[1]Akdeniz University, Faculty of Education, Antalya, Türkiye
· **piert34@gmail.com** · ORCID > 0000-0002-3785-2384

[2]Hacettepe University, Faculty of Education, Ankara, Türkiye
· **hulyaebb@hacettepe.edu.tr** · ORCID > 0000-0002-0741-9934

Makale Bilgisi/Article Information

Makale Türü/Article Types: Araştırma Makalesi/Research Article
Geliş Tarihi/Received: 14 Ocak/January 2024
Kabul Tarihi/Accepted: 05 Şubat/February 2024
Yıl/Year: 2024 | **Cilt-Volume:** 43 | **Sayı-Issue:** 2 | **Sayfa/Pages:** 951-970

Atıf/Cite as: Ukşul, E. & Kelecioğlu. H. "Investigation of Models Used in Equating Testlet-Based Tests"
Ondokuz Mayıs University Journal of Faculty of Education,
43(2), December 2024: 951-970.

Sorumlu Yazar/Corresponding Author: Ertunç UKŞUL

Etik Kurul Beyanı/Ethics Committee Approv: "Ethics committee approval was received for the research from the Scientific Research and Publication Ethics Committee of Hacettepe University with the decision number 35853172-300 dated 21.04.2020."

https://doi.org/10.7822/omuefd.1419482

INVESTIGATION OF MODELS USED IN EQUATING TESTLET-BASED TESTS

ABSTRACT

This study aims to examine the effects of testlets on test equating. For this purpose unidimensional item response theory, two-factor item response theory and testlet response theory models were applied to the testlet-based tests for the estimation of item and ability parameters. In order to equate the tests, the parameters were placed on the common scale using mean-mean, mean-sigma and Stocking-Lord scale transformation methods under the common-item non-equivalent groups design. Then, the equating errors of the models depending on the scale transformation method and the number of testlets were calculated and compared. Equating errors were compared with Root Mean Squared Error. In the study, the science test of the Trends in International Mathematics and Science Study project administered in 2019 was used as the data collection tool. As a result of the study, it was determined that the use of unidimensional item response theory model increased the equating error, while the use of two-factor and testlet response theory models decreased the equating error as the number of testlets in the test increased. To compare the models, the correlation between the parameters obtained from the models after scale transformation was examined and it was found that the item parameters were more affected by the model selection than the ability parameter. In addition, it was concluded that the equating errors obtained from the mean-mean and Stocking-Lord scale transformation methods were lower than the mean-sigma method.

Keywords: Test Equating, Scale Transformation, Item Response Theory, Testlet Response Model, Two-factor Model.

MADDE TAKIMI TABANLI TESTLERİN EŞİTLENMESİNDE KULLANILAN MODELLERİN İNCELENMESİ

ÖZ

Bu çalışmada, testlerde yer alan madde takımlarının uygulanan modellere göre test eşitleme hatasına olan etkisi araştırılmıştır. Bu amaçla, çalışma kapsamında analiz edilen madde takımı içeren testlere madde tepki kuramı modellerinden tek boyutlu madde tepki kuramı, iki faktör madde tepki kuramı ve madde takımı tepki kuramı modelleri uygulanarak madde ve yetenek parametreleri kestirilmiştir. Testlerin eşitlenebilmesi için elde edilen parametreler, eşdeğer olmayan gruplar ortak madde deseni altında ortalama-ortalama, ortalama-sigma ve Stocking-Lord ölçek dönüştürme yöntemleri kullanılarak ortak ölçeğe yerleştirilmiştir. Daha sonra modellerin ölçek dönüştürme yöntemine ve madde takım sayısına bağlı olarak değişen eşitleme hataları hesaplanmış ve karşılaştırılmıştır. Değerlendirme ölçütü olarak Root Mean Squared Error tercih edilmiştir. Araştırmada, veri toplama aracı olarak Trends in International Mathematics and Science Study projesinin 2019 yılında uygulanan fen bilimleri testi kullanılmıştır. Araştırma sonucunda, testte yer alan madde takım sayısı arttıkça tek boyutlu madde tepki kuramı modeli kullanımının eşitleme hatasını artırdığı, iki faktör ve madde takımı tepki kuramı modelleri kullanımının ise eşitleme hatasını düşürdüğü belirlenmiştir. Modellerin karşılaştırılması için ölçek dönüştürme sonrası modellerden elde edilen parametreler arasındaki ilişki incelenmiş, madde parametrelerinin yetenek parametresine göre model seçiminden daha fazla etkilendiği bulunmuştur. Bununla birlikte ortalama-ortalama ve Stocking-Lord ölçek dönüştürme yöntemlerinden elde edilen eşitleme hatalarının ortalama-sigma yöntemine göre daha düşük olduğu sonucuna ulaşılmıştır.

Anahtar Sözcükler: Test Eşitleme, Ölçek Dönüştürme, Madde Tepki Kuramı, Madde Takımı Tepki Modeli, İki Faktör Model.

INTRODUCTION

The educational system is a crucial structure that contributes to the development of individuals and societies. One of the fundamental elements shaping the educational system is the use of examinations and assessment tools designed to evaluate students’ knowledge, skills, and abilities. The valid and reliable execution of these assessments holds great significance in steering the educational system in the right direction.

Large-scale exams are commonly employed in the assessment of education systems across countries. These comprehensive tests are required to be administered to different groups at various times. The questions and test booklets used in these assessments often exhibit variations within themselves. Therefore, student performance data obtained at different times or through different measurement tools for the same purpose must be comparable because discrepancies in administered tests can lead to errors in the evaluation and comparison of tests and students. Test equating is necessary to mitigate these errors, achieve valid and reliable results from tests, and interpret the obtained results accurately (Cook & Eignor, 1991). Consequently, test equating becomes a significant process for large-scale exams where multiple test booklets are used in each administration (Gübeş & Kelecioğlu, 2015).

The objective of test equating is to adjust the score scales related to different test forms and compensate for the relative difficulty variations stemming from the measurement instrument. In this way, test scores are rendered equivalent and comparable. Through this process, the obtained scores from different test forms designed for the same purpose can be interchangeably used (Angoff, 1984; Hambleton & Swaminathan, 1985). The opportunity provided by test equating helps to overcome the advantages or disadvantages resulting from differences in difficulty levels among test forms, reducing the negative effects arising from students taking an easier or more challenging test form (Tan, 2005).

According to Cook & Eignor (1991), one of the main purposes of a standardized test is to provide a fair and equitable psychological or educational assessment. Test equating allows student performance to be monitored over time and achievement differences between different groups of students to be analyzed in a standardized way. This makes it possible to assess and compare students fairly.

For test equating, it is imperative to place the test scores obtained from different test forms onto a common scale. Various equating processes provide transformations of scores related to test forms. Some of these processes use Item Response Theory (IRT) methods, while others employ Classical Test Theory (CTT) methods.

CTT methods such as mean, linear and equipercentile methods, provide information primarily at the test level without constructing complex theoretical models and rely on a weak mathematical foundation to estimate item parameters based on group-related data. In contrast, IRT methods define the relationship between an individual's performance and responses to items through a function called the "item characteristic curve" and have a strong mathematical background that estimates item and test statistics independently from the group and ability levels independently from the test (Hambleton & Swaminathan, 1985). Since the parameters estimated by IRT are not dependent on the group, they have the property of invariance and it is possible to compare different groups with each other based on the results obtained by applying these methods (Embretson & Reise, 2000).

As Bobcock (2009) points out, IRT overcomes the limitations of CTT and develops stronger assumptions. In overcoming these limitations, IRT relies on a strong mathematical foundation and involves complex algorithms that can be processed by computers. Therefore, it is very important to ensure the necessary assumptions and select the right mathematical models for IRT to provide reliable results. The model-data fit of the applied model directly affects the algorithm used to calibrate the items and thus the accuracy of the equating results (Hambleton & Swaminathan, 1985; Zhao, 2008).

Certain assumptions need to be met to be able to equate with the IRT. One of the most important of these assumptions is local independence. Local independence means that the responses to any two items are statistically independent of each other when the ability variable is held constant (Lord, 1980). In other words, the probability of answering items correctly does not depend on any factor other than the ability of the individual at a given ability level (Hambleton & Swaminathan, 1985). If the content of an item contains clues or directions about the answer of another item and affects the correct answer of that item, local independence is violated in that test (Embretson & Reise, 2000).

There are many different causes and types of local item dependency (LID). One of the most important is passage dependency. Many large-scale tests include a question type called testlet which consists of multiple question items linked to the same passage which allows for complex and interrelated questions and makes the test more time-efficient (DeMars, 2006). Due to this content of large-scale tests, studies on local dependence have mostly focused on passage dependence, and it has been determined that the interrelatedness of testlets and their effects on finding the correct answer cause local dependence (Koğar & Kelecioğlu, 2017). Therefore, this study focuses on the testlets that cause local item dependence in the test equating process.

Although the use of testlets in tests dates back to earlier times, the definition of a testlet was first described by Wainer and Kiely in 1987 (Wainer et al., 2007). Since then, the effects of testlets on tests have been considered remarkable by educational statisticians and various studies have been conducted on this subject. These studies have shown that the application of traditional IRT models that ignore the effect of testlets on tests violates the local independence assumption of IRT and leads to equating/scaling errors. For this reason, an alternative theory called Testlet Response Theory (TRT) was developed as an alternative model to unidimensional item response theory (UIRT) (Lee et al., 2001).

TRT model developed by Wainer et al. (2000) is based on the two-factor model (2FM), one of the multidimensional item response theory (MIRT) models, and treats the testlets that constitute local dependence as a secondary dimension to be measured. For this purpose, the random effect parameter, which shows the effect of the shared variance between the items in the testlets, is added to the traditional UIRT parameters.

The traditional 2PL UIRT equation is shown in Equation (1), while the 2PL TRT equation where the testlet effect (γg(i)) is added to the 2PL UIRT is shown in Equation (2).

$$P_i(\theta) = \frac{e^{(a_i\ (\theta - b_i)}}{1 + e^{(a_i\ (\theta - b_i)'}} \tag{1}$$

$$P_i(\theta) = \frac{e^{(a_i\ (\theta - b_i - \gamma_{g(i)}))}}{1 + e^{(a_i\ (\theta - b_i - \gamma_{g(i)}))}} \tag{2}$$

P(·): probability that person j responds correctly to item *I*

a_i: the discrimination parameter of item *i*

θ *:* the ability level of person *j*

b_i: item difficulty parameter of item *i*

$\gamma_{g(i)}$: the random testlet effect for person j of the testlet *g(i)* to which item *i* belongs.

As the variance for $\gamma_{g(i)}$ increases, a larger effect is observed for testlet *g* (Bradlow et al., 1999).

TRT is a constrained version of the 2FM. In the 2FM, similar to the TRT model, items are regarded as conditionally independent among testlets; however, within testlets, they are deemed conditionally dependent. Nevertheless, the primary distinction between the 2FM and TRT model lies in the fact that the 2FM permits separate discrimination parameters for the primary and secondary (testlet) dimensions, and these discrimination parameters are capable of operating independently from one another (Bao, 2007; Rijmen, 2009; Wainer et al., 2007).

According to Kolen & Brennan (2014), the initial phase in the equating process is to decide on the equating design required for data collection. Single group design, common item in non-equivalent groups design and equivalent groups design are the most commonly used equating designs. Since this study was conducted on TIMSS 2019 data, the common item in non-equivalent groups design was used in which the differences between groups were controlled by common items (anchor) in each booklet and the test booklets were equated with each other through these common items. The common item design in non-equivalent groups is shown in Figure 1.

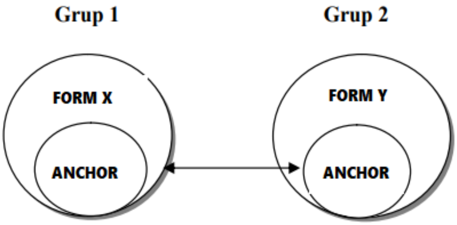

Figure 1. *Common Item Pattern in Non-Equivalent Groups*

After selecting the appropriate pattern for the data in equating, the equating method is decided. Equating methods is the process used to perform test equating (Öztürk, 2010). In this process, item parameters are estimated according to the appropriate method. Equating methods are categorized under two main headings: methods based on CTT and methods based on IRT. Since methods based on CTT contain less information and more errors on item basis (Hambleton, 1985), in this study, equating methods based on UIRT, 2FM and TRT were used instead of CTT.

According to Kolen & Brennan (2014), after the item and ability parameters are estimated with the appropriate method if a common item design is to be used in non-equivalent groups, the estimated parameters should be placed on a common scale to make them comparable. Since the parameters are obtained from different samples, they are on different scales. This process of placing the estimated parameters on the same scale through a linear transformation is called "calibration". In the calibration process, there are two different calibration methods: separate calibration and concurrent calibration. In this study, the separate calibration method was used because the separate calibration method gives more accurate and reliable results in multidimensional data (Kolen & Brennan, 2014).

In the separate calibration method, the estimated item parameters should be placed on the same scale before test equating. In scale transformation, the item and ability parameters estimated from the new form are converted to the scale of the old form through the parameters of the common items in the old form.

For the transformation, the slope (A) and intercept (B) constants are first obtained as in Equation (3).

$$a_{Ei} = \frac{a_{Yi}}{A}$$

$$b_{Ei} = Ab_{Yi} + B \tag{3}$$

a_{Ei}: discrimination parameter of item *i* obtained from test E (old form)

b_{Ei}: difficulty parameter of item *i* obtained from test E (old form)

a_{Yi}: rescaled discrimination parameter of item *i* obtained from test *Y* (new form)

b_{Yi}: rescaled difficulty parameter of item *i* obtained from test *Y* (new form)

Using the constants A and B obtained from Equation (3), the equivalent of the ability level (θ) of person *j* in test *Y* is calculated as in Equation (4).

$$\theta_{Ej} = A\theta_{Yj} + B \tag{4}$$

The methods used for scale transformation during separate calibration are divided into two: moment methods are mean-mean (MM) and mean-sigma (MS) methods; characteristic curve methods are Stocking-Lord (SL) and Haebara (HA) methods. The difference between these methods is the difference in the methods of obtaining the A and B coefficients required for scale transformation (Embretson & Reise, 2000). In this study, MM, MS and SL methods were used.

The MM method introduced by Loyd & Hoover (1980) considers the item discriminations and item difficulties and uses Equation (5) to estimate the scaling constants A and B as follows:

$$A_{MM} = \frac{M(a_N)}{M(a_O)}$$

$$B_{MM} = M(b_O) - A_{MM}M(b_N) \tag{5}$$

$M(\cdot)$: operator for the arithmetic mean

a_N, b_N: item discrimination, item difficulty parameters on the new scale

a_O, b_O: counterpart on the old scale

The MS method, proposed by Marco (1977), considers the item difficulties and uses Equation (6) to estimate the scaling constants A and B as follows;

$$A_{MS} = \frac{SD(b_O)}{SD(b_N)}$$

$$B_{MS} = M(b_O) - A_{MS}M(b_N) \tag{6}$$

$SD(\cdot)$: operator for the standard deviation

b_N: item difficulty parameter on the new scale

b_O: counterpart on the old scale

The SL method, proposed by Stocking & Lord (1983), articulated the difference between characteristic curves by squaring the sum of the differences between item characteristic curves for each item within a specific θ. Expressing this for a given θ_i, the summation of squared differences over anchor items (*j*:*V*) can be represented using Equation (7) as outlined by Kolen and Brennan (2004).

$$SLdiff(\theta_i) = \left[\sum_{j:V} p'_{ij}(\theta_{Ji}, a_{Jj}, b_{Jj}, c_{Jj}) - \sum_{j:V} p_{ij}(\theta_{Ji}, \frac{a_{Ij}}{A}, Ab_{Ij} + B, c_{Ij}) \right]^2$$

$$SL_{crit} = \sum_{i} SLdiff(\theta_i) \tag{7}$$

A : Slope constant

B : Intercept constant

P_{ij} (·): Item characteristic function

P'_{ij} (·): Equated item characteristic function

To obtain equating errors, slope (A) and intercept (B) coefficients were calculated separately with MM, MS and SL methods, then the Root Mean Squared Error (RMSE) value, which gives the amount of error related to the item parameters estimation, was obtained.

The RMSE formula for a sample is shown in Equation (8).

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}\left(\tau_{jr} - \tau_j\right)^2}{R}} \tag{8}$$

j : actual value of parameter j

jr : predicted value of parameter j for the rth observation

R : number of observations

This study aims to compare the RMSE values obtained from UIRT, 2FM and TRT models when the number of testlets changed according to MM, MS and SL scaling methods. In this research, the models utilized for equating testlet-based tests are compared by employing real data under conditions such as the number of item bundles, the number of independent items and the sample size. Therefore, it is believed that the results of the research will provide significant contributions to the relevant literature.

The research question that shapes the study is as follows:

How does the error of the scale transformation methods vary according to the number of testlets when unidimensional item response theory, two-factor models, and item bundle response theory models are used in testlet-based tests?

METHOD

Research Design

This study aims to investigate the variation of scale transformation errors according to the IRT models. The study is a basic research study since it aims to perform model comparison and test a developed theory.

Participants

The data were obtained from the science test of the "TIMSS (Trends in International Mathematics and Science Study) project conducted in 2019 by the International Association for the Evaluation of Educational Achievement (IEA)" (Richardson et al., 2020)

TIMSS is a worldwide initiative conducted in many countries and repeated every four years, aimed at assessing the knowledge and skills students acquire in mathematics and science. The TIMSS student population comprises students in the 4th and 8th grades.

In TIMSS 2019, approximately 250,000 students from 8th grade and about 310,000 students from 4th grade participated across 39 countries. The study featured 14 question booklets, among which booklets 7 and 8 were used as the data collection instruments for equating purposes. While selecting the data, no country distinction was made and all respondents who took these two booklets and whose data were available were included in the study. After data cleaning processes, 7,988 students who received booklet 7 and 7,946 students who received booklet 8 formed the sample of the study.

Research Data

In TIMSS 2019, there are 14 different booklets containing the science cognitive test. In this study, a purposive sampling method was used and Booklet 7 was selected as old form and Booklet 8 was selected as new form for test equating because they contain many common items and testlets.

In the 7th booklet (old form), a total of 45 items, including 22 independent items and 4 testlets, were included in the analysis. The 4 testlets in the booklet consist of 4, 8, 3 and 6 items respectively. In the 8th booklet (new form), a total of 42 items, including 22 independent items and 4 testlets, were analyzed. The 4 testlets in the booklet consisted of 4, 8, 3 and 5 items, respectively.

The first 26 items of the 7th and 8th booklets, including the first 11 independent items and the first 3 testlets (testlets consisting of 3, 4 and 8 items), are common items. In the common item non-equivalent groups design, the equations required for the scale transformation process are obtained over the common items.

Four separate data sets were created using the testlets in the 7th and 8th booklets. The number of independent items and testlets in the data sets are given in Table 1.

Table 1. *Number of Testlets and Independent Items in the Data Sets Used*

		Set 1	Set 2	Set 3	Set 4
Booklet 7 (old form)	**Number of Testlets (Number of Items)**	1 (4)	2 (4,8)	3 (4,8,3)	4 (4,8,3,6)
	Independent Items	24	24	24	24
	Total	28	36	39	45
Booklet 8 (new form)	**Number of Testlets (Number of Items)**	1 (4)	2 (4,8)	3 (4,8,3)	4 (4,8,3,5)
	Independent Items	22	22	22	22
	Total	26	34	37	42
Common Item (anchor)	**Number of Testlets (Number of Items)**	1 (4)	2 (4,8)	3 (4,8,3)	3 (4,8,3)
	Independent Items	11	11	11	11
	Total	15	23	26	26

As seen in the Table 1, all data sets created from Booklet 7 contains 24 independent items and Set 1 contains 1 testlet (with 4 items), Set 2 contains 2 testlets (with 4 and 8 items), Set 3 contains 3 testlet (with 4, 8, 3 items), Set 4 contains 4 testlets (with 4, 8, 3, 6 items) respectively. All data sets created from Booklet 8 contains 22 independent items and Set 1 contains 1 testlet (with 4 items), Set 2 contains 2 testlets (with 4 and 8 items), Set 3 contains 3 testlet (with 4, 8, 3 items), Set 4 contains 4 testlets (with 4, 8, 3, 5 items) respectively. All data sets created from Common Items contains 11 independent items and Set 1 contains 1 testlet (with 4 items), Set 2 contains 2 testlets (with 4 and 8 items), Set 3 contains 3 testlet (with 4, 8, 3 items), Set 4 contains 3 testlets (with 4, 8, 3 items) respectively.

The item and ability parameters examined in the study were estimated using the 2PL models of UIRT, 2FM and TRT. In this way, it was aimed to determine the effect of the number of testlets on the equating error according to the models and methods used.

Data Analysis

In this study, item and ability parameters were estimated from Booklet 7 and Booklet 8 of science cognitive test in TIMSS 2019. These parameters were obtained using the UIRT, 2FM and TRT models in an item calibration and test scoring application using item response theory. In the data sets, a (slope), b (slope-threshold), c (slope-intercept) and theta (ability) parameters were obtained separately for all items. Since the b (slope-threshold) parameter in multidimensional models is not interpreted similarly to the unidimensional models and cannot make accurate generalizations, it is more appropriate to use the c (b=-c/a) parameter instead (Cai et al., 2011; Min & He, 2014). For this reason, the c parameter was used in all models to make comparisons between models.

Item parameter estimation is performed with the Bock-Aitkin marginal maximum likelihood estimation method (Bock & Aitkin, 1981) and ability parameter estimation is performed with the expected a posteriori (EAP) method since these methods can make more effective estimations for unidimensional and two-factor IRT models (Cai et al., 2011).

In order to display the obtained parameters on the same scale, scale transformation was applied to the parameters of the items in the 8th booklet based on the parameters of the common items in the 7th booklet. For the scale transformation, equating software was used for unidimensional models and multidimensional models, and mean-mean (MM), mean-sigma (MS) and Stocking Lord (SL) separate calibration scale transformation methods were applied to the models.

Local dependence *X*2 (LD X^2) statistic was used to analyse whether the items in the testlets were locally dependent due to its ease of calculation and excellent performance (Liu & Thissen, 2012) . Values of LD X^2 greater than 10 indicate high local dependence between items, values 5-10 indicate moderate local dependence, and values less than 5 indicate low local dependence (Cai et al., 2011). The LD X^2 ratio values of the items in the testlets in the study are given in Table 2. The testlets coded with 1, 2 and 3 in Table 2 are included in both booklets, namely, they contain common items. Testlets coded with 4 and 5 represent the testlets consisting of different questions in the two booklets.

Table 2. *LD X^2 Values of the Items in the Testlets*

Booklets	Testlets	LD X^2 *Range*
Booklet 7	Testlet 1	(31,9 - 330,7)
	Testlet 2	(-0,5 - 504,4)
	Testlet 3	(11,3 - 41,1)
	Testlet 4	(0,0 - 107,0)
Booklet 8	Testlet 1	(11,4 - 229,2)
	Testlet 2	(0,4 - 475,4)
	Testlet 3	(-0,4 - 29,3)
	Testlet 5	(1,8 - 323,3)

As Table 2 is examined, LD X^2 Range of the testlets in Booklet 7 is between -0,5 and 504,4; the testlets in Booklet 8 is between -0,4 and 475,4. The LD X^2 values of the item pairs in the testlets are generally above 10 and at a high level. Based on this finding, it can be said that there is local dependency between item pairs and that the use of 2FM and TRT methods would be appropriate for both booklets.

Ethics Committee Approval

Ethics committee approval was received for this study from Hacettepe University, Faculty of Education.

The Title of The Ethics Committee: Scientific Research and Publication Ethics Committee of Hacettepe University

Approval Date: 21.04.2020,

Ethics Document's Number: 35853172-300

FINDINGS

Scale transformation methods were applied to the data obtained in UIRT, 2FM and TRT models. The RMSE values for the item and ability parameters related to the general dimension obtained through the process of scale transformation are presented in Table 3.

Table 3. *Scale Transformation RMSE Values of Item and Ability Parameters*

	Scale Conversion		RMSE		
Data Set	**Method**	**Model**	**a**	**c**	**theta**
Data Set 1 (1 testlets)	OO	UIRT	0,0394	0,0698	0,0466
		2FM	0,0267	0,1045	0,0882
		TRT	0,0272	0,1071	0,0904
	OS	UIRT	0,0159	0,0622	0,0723
		2FM	0,0532	0,1123	0,0739
		TRT	0,0673	0,1237	0,0731
	SL	UIRT	0,0045	0,0569	0,0600
		2FM	0,0267	0,0861	0,0723
		TRT	0,0272	0,0888	0,0723
Data Set 2 (2 testlets)	OO	UIRT	0,0713	0,1101	0,0723
		2FM	0,0974	0,1687	0,0754
		TRT	0,0545	0,1534	0,0805
	OS	UIRT	0,0157	0,0835	0,0723
		2FM	0,0154	0,1109	0,0927
		TRT	0,0545	0,1534	0,0805
	SL	UIRT	0,0032	0,0501	0,0500
		2FM	0,0202	0,0795	0,0532
		TRT	0,0302	0,1303	0,0940
Data Set 3 (3 testlets)	OO	UIRT	0,0917	0,1441	0,0789
		2FM	0,0840	0,1564	0,0606
		TRT	0,0386	0,1553	0,1025
	OS	UIRT	0,0119	0,0928	0,0796
		2FM	0,0330	0,1198	0,0742
		TRT	0,0583	0,1629	0,0804
	SL	UIRT	0,0551	0,1159	0,0546
		2FM	0,0072	0,0506	0,0410
		TRT	0,0148	0,1177	0,1004

Data Set 4 **(4 testlets)**	OO	UIRT	0,0785	0,2616	0,3132
		2FM	0,0746	0,1447	0,0563
		TRT	0,0399	0,1557	0,1050
	OS	UIRT	0,0816	0,1710	0,0856
		2FM	0,0357	0,1222	0,0761
		TRT	0,0591	0,1613	0,0832
	SL	UIRT	0,0223	0,1376	0,1511
		2FM	0,0128	0,0556	0,0410
		TRT	0,0055	0,1162	0,1104

As seen in Table 3, the RMSE value increases in UIRT as the number of testlets in the data set increases, while the RMSE value decreases in 2FM and TRT. In Dataset 1, which contains the least number of testlet, the RMSE values of a, c and theta parameters obtained from UIRT are lower than the other models, while in Dataset 4, which contains the highest number of testlets, the RMSE values obtained from 2FM and TRT are lower. While the lowest RMSE values in TRT and 2FM are obtained in Dataset 4, which contains the most testlets, the lowest RMSE values in UIRT are obtained in Dataset 1, which contains the fewest testlets.

When the RMSE values obtained for the methods are analysed, it is seen that the RMSE values for the mean-sigma method are generally higher than the other methods. In addition, it is seen that the lowest RMSE values are obtained from the Stocking Lord method.

The RMSE values obtained for a, c and theta parameters as a result of the scale conversion process are shown graphically in Figure 2.

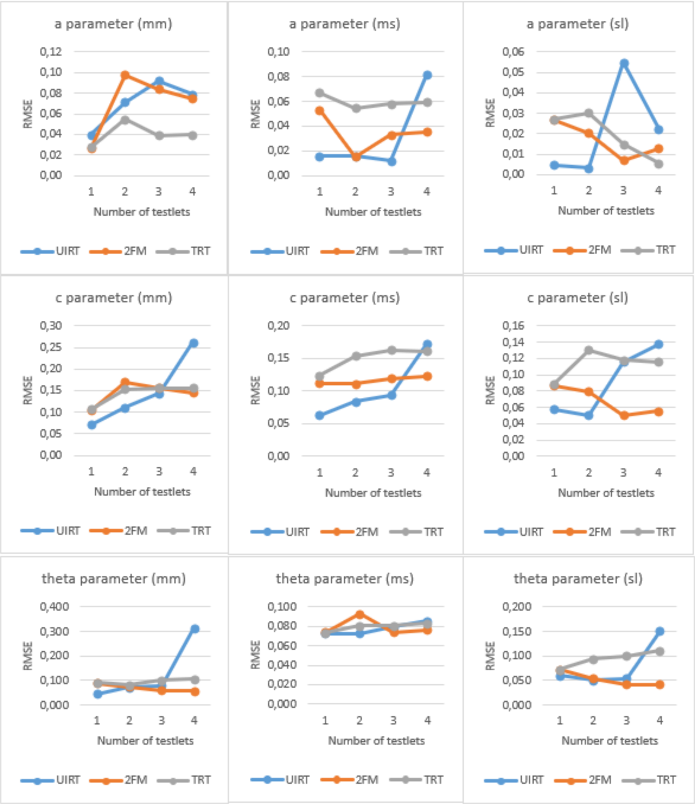

Figure 2. *Sale Conversion RMSE Values of Item and Ability Parameters under All Conditions*

When Figure 2 is examined, it is seen that as the number of testlets in the test form increases, the RMSE values obtained from UIRT generally increase, while the RMSE values obtained from 2FM and TRT generally decrease. Although the RMSEs obtained from 2FM and TRT show a similar pattern, it is seen that 2FM has generally lower RMSE values than TRT in tests with a high number of testlets. When the scale transformation methods are analyzed, it is seen that as the number of testlets increases, the lowest errors in the mean-mean method are obtained in the TRT, and the lowest errors in the mean-sigma and Stocking-Lord methods are obtained in the 2FM model.

In order to analyse the relationship between the models after scale transformation, the correlation values of the models with each other were analysed. The findings are shown in Table 4.

Table 4. *Correlation of Item and Ability Parameters after Scale Transformation*

Data Set	TRT-2FM			TRT-UIRT			2FM- UIRT		
	a	c	theta	a	c	theta	a	c	theta
1	0,9894	0,9963	0,9997	0,9372	0,9912	0,9926	0,9140	0,9816	0,9930
2	0,9762	0,9962	0,9968	0,9183	0,9858	0,9932	0,8979	0,9784	0,9939
3	0,9541	0,9952	0,9972	0,9219	0,9860	0,9943	0,8782	0,9774	0,9948
4	0,9579	0,9954	0,9972	0,9238	0,9859	0,9935	0,8866	0,9780	0,9935

As Table 4 is examined, it is seen that the correlation between the values of a, c and theta parameters obtained from three different models as a result of scale transformation are generally high. The correlation of TRT-2FM methods is the highest, while the correlation of 2FM-UIRT methods is the lowest.

Especially as the number of items in the data set increases, the correlation between TRT-UIRT and 2FM-UIRT decreases. Based on this, it can be said that 2FM and TRT methods diverge positively from UIRT as the number of testlets in the test increases.

When the correlation of theta parameters is analysed, it can be seen that the highest correlation is obtained in the "theta" parameter. All three models have similar findings on "theta" parameter. If the correlation of "a" parameters is analyzed, it can be seen that the lowest correlation is obtained on "a" parameter. Accordingly, it can be said that the methods used affect the "a" parameter more than the other parameters and theta parameter less than the others.

DISCUSSION, CONCLUSION AND SUGGESTIONS

In this study, scale transformation methods were applied to two booklets (booklets 7 and 8) of the TIMSS 2019 science test under the common-item non-equivalent groups design using UIRT, 2FM and TRT models, and the equating errors (RMSE) obtained from different scaling methods were compared as the number of testlets in the test increased.

When the data of the study was analysed, it was seen that the local dependency values in the testlets were generally high (LD X^2 >10). Various studies have shown that the application of traditional IRT models that ignore the effect of testlets violates the local independence assumption of IRT and leads to equating/scaling errors (Lee et al., 2001). In this study it would be appropriate to use 2FM and TRT models with UIRT due to high local dependence in the testlets.

In the literature, there are numerous studies indicating that the violation of the local independence assumption creates issues in item parameter estimations and error calculations associated with item parameters when using standard IRT models (Ackerman, 1987; Wainer, 1995; Wainer & Wang, 2000). In this study, it has been determined that the equating error of discrimination parameter (a), difficulty parameter (c) and ability parameter (θ) obtained from the UIRT, increase with the number of testlets.

According to the research findings as the number of testlets decreased the UIRT model provided lower equating errors than 2FM and TRT models however the number of testlets increased, 2FM and TRT models provided lower equating errors than the UIRT model. The error values obtained from 2FM and TRT models have similar patterns in general. As a result of the study, it was observed that the number of testlets had an negative effect on the equating errors in the UIRT model. There are similar studies in the literature recommending the use of 2FM and TRT models in testlet-based tests (Bradlow & Wainer, 2002, Demars, 2006; Wang).

It was observed that "a" and "c" parameters were more sensitive to the model used, while the ability parameter (θ) showed a similar pattern in all three models. The findings of this study, showing high correlations among errors obtained from different models of general ability, align with previous studies in the field. DeMars (2006) and Bradlow et al. (1999) in their respective studies, expressed similarity in ability parameters. Especially in studies examining "a" and "c" parameters, it may be recommended to use 2FM and TRT models instead of the UIRT model as the number of testlets increases. In studies examining the "θ" parameter, all three models can be used interchangeably under certain conditions.

In this study using the mean-sigma method in equating studies would lead to higher equating error. The use of mean-mean and Stocking-Lord methods in the studies may affect on reducing equating errors. Baker & Al-Karni (1991), Gök & Kelecioğlu (2014) and Zor (2023) in their respective studies, have obtained similar results. In the study conducted by Zor and Gök & Kelecioğlu, the mean-mean method yielded the lowest equating error, while in the study by Baker & Al-Karni, the Stocking-Lord method provided the lowest equating error.

This study aimed to compare the equating errors (RMSE) of parameters obtained from various IRT models and scale transformation methods when the number of testlets changed. As examining the results, it can be stated that choosing the correct model and scale transformation method will reduce equating errors. Since actual data were used in this study, it is thought that the results of the study will be important for the related literature.

The results obtained in the study were obtained from actual data. ~~While~~ Working with actual data may cause higher errors than simulation studies. A similar study can be examined with simulation data and the results can be compared.

The effects of the models and methods used in equating can be examined and compared in smaller or large samples. In his study, Zhang (2010) indicated that the sample size had an impact on the outcomes for all UIRT, 2FM and TRT models, with better results obtained from larger samples. In this study, actual data were utilized and the data were obtained from the actual responses of 7,946 participants. It can be stated that a large sample was used in this study. Similar studies can be examined with lower samples and the results can be compared.

The actual data used in this study were obtained from two booklets containing 42 and 46 items, respectively. In his simulation study DeMars (2006) computed the average RMSE values for primary ability scores using 2FM and TRT models for two datasets consisting of 25 and 50 items and found that, for all models, the RMSE values decreased as the test length increased. Similar studies can be conducted with tests containing fewer or more items.

In this study, 2PLM was used for all models. In future studies, the errors obtained from the analysis with 3PLM can be compared.

Since actual data were used in the study, the number of testlets was limited to 4 due to the test. Equating errors of tests with more testlets can be examined with actual data or simulation studies with more testlets.

RMSE value was used as an evaluation criterion in this study. In other studies, equating errors can be compared using different evaluation criteria.

CONFLICT OF INTEREST

There is no personal or financial conflict of interest between the authors of the article within the scope of the study.

AUTHOR CONTRIBUTIONS

Research design: EU(%50), HK(%50)

Data collection: EU(%50), HK(%50)

Statistical analysis: EU(%50), HK(%50)

Preparation of the Article: EU(%50), HK(%50)

REFERENCES

Ackerman, T. A. (1987). *ACT research report series: The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. PsycEXTRA Dataset. https://doi.org/10.1037/e426132008-001

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Babcock, B. G. E. (2009). *Estimating a Noncompensatory IRT Model Using a modified Metropolis algorithm* [Unpublished doctoral dissertation]. The University of Minnesota.

Baker, F. B. & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147-162. https://doi.org/10.1111/j.1745-3984.1991.tb00350.x

Bao, H. (2007). *Investigating differential item function amplification and cancellation in the application of item response testlet models* [Unpublished doctoral dissertation]. University of Maryland.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/bf02293801

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168. https://doi.org/10.1007/bf02294533

Cook, L. L. & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00207.x

DeMars, C. E. (2006). Application of the Bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Gök, B. & Kelecioğlu, H. (2014). Comparison of IRT Equating Methods Using the Common-Item Nonequivalent Groups Design. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136.

Gübeş, N. Ö. & Kelecioğlu, H. (2015). Karma Testlerin Eşitlenmesinde MTK Eşitleme Yöntemlerinin Eşitlik Özelliği Korunumu Ölçütüne Göre Karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1). https://doi.org/10.21031/epod.65039

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer Science & Business Media.

Kogar, E. Y. & Kelecioglu, H. (2017). Examination of different item response theory models on tests composed of Testlets. *Journal of Education and Learning, 6*(4), 113. https://doi.org/10.5539/jel.v6n4p113

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer Science & Business Media.

Lee, G., Kolen, M. J., Frisbie, D. A. & Ankenmann, R. D. (2001). Comparison of dichotomous and Polytomous item response models in equating scores from tests composed of Testlets. *Applied Psychological Measurement, 25*(4), 357-372. https://doi.org/10.1177/01466210122032226

Liu, Y. & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the Bifactor logistic model. *Applied Psychological Measurement, 36*(8), 670-688. https://doi.org/10.1177/0146621612458174

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4), 453-477. https://doi.org/10.1177/0265532214527277

Richardson, M., Isaacs, T., Barnes, I., Swensson, C., Wilkinson, D. & Golding, J. (2020). *Trends in International Mathematics and Science Study (TIMSS) 2019: National Report for England*. Research Report. UK Department for Education.

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. *ETS Research Report Series, 2009*(2). https://doi.org/10.1002/j.2333-8504.2009.tb02194.x

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. https://doi.org/10.1177/014662168300700208

Tan, Ş. (2005). Küçük örneklemlerde beta4 ve polynomial loglineer öndüzgünleştirme ve kübik eğri sondüzgünleştirme metotlarının uygunluğu. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 35*(1), 123-151.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8*(2), 157-186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H. & Wang, X. (2000). Using a new statistical model for Testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

Zhao, Y. (2008). *Approaches for addressing the fit of item response theory models to educational test data* [Unpublished doctoral dissertation] University of Massachusetts Amherst.

Zor, Y. M. (2023). ***Çok boyutlu testlerin tek boyutlu ve çok boyutlu yöntemlere göre eşitlenmesi*** [Unpublished doctoral dissertation] Hacettepe University.

❁ ❁ ❁